Generalization Bound

Tri Nguyen

nguyetr9@oregonstate.edu

May 11, 2023

Though I got it but ...

Generalization bound is a characterization of a function class, measuring how hard the function class can be *learned* using a finite sample. Let $L_S(f), L_{\mathcal{D}}(f)$ be empirical loss and true loss of a predictor f. Generalization bound is defined as

$$|L_{\mathcal{D}}(f) - L_S(f)|$$

We wish this bound to be small. It depends on size of dataset S. We have not said anything about the relationship between f and S.

At a glance, if f is some given fixed predictor, then the bound can be bounded by concentration inequality. Let us be clear by defining these losses.

$$L_S(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, \boldsymbol{z}_i), \quad \boldsymbol{z}_1, \ldots, \boldsymbol{z}_i \sim_{\text{i.i.d.}} \mathcal{D}$$
$$L_{\mathcal{D}}(f) = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{D}} \left[\ell(f, \boldsymbol{z})\right],$$

where $\ell(f, \boldsymbol{z})$ is the loss evaluated at data point $\boldsymbol{z}$. Now if we also assume that $0 \leq \ell(f, \boldsymbol{z}) \leq C$, then we revoke typical concentration inequality like Hoeffding inequality.

Theorem 1 (Hoeffding inequality)**.** *Let $X_1, \ldots, X_n$ be independent, and $a_i \leq X_i \leq b_i, i \in [n]$, $S_n \triangleq \sum_{i=1}^n X_i$. The for $t > 0$,*

$$\mathsf{Pr}(S_n - \mathbb{E}[S_n] \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Calling Theorem 1 we get

$$\mathsf{Pr}(nL_S(f) - nL_{\mathcal{D}}(f) \geq t) \leq \exp\left(-\frac{2t^2}{nC^2}\right)$$
$$\mathsf{Pr}(L_S(f) - L_{\mathcal{D}}(f) \geq \frac{t}{n}) \leq \exp\left(-\frac{2t^2}{nC^2}\right)$$

Equivalently,

$$\mathsf{Pr}\left(L_S(f) - L_{\mathcal{D}}(f) \geq C\sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta \tag{1}$$

the above derivation should be made with absolute operator.

The bound looks very standard. The pitfall here is that f **is independent to the dataset** S, which is not true. The key is that the randomness is from the data S. One way to interpret the claim in (1) is: Given a fixed predictor f, we draw randomly data set S 100 times, then there would be not more than 100δ times that the generalization error is large. Then it is apparent that (1) is not applicable to predictors f that depend on data S, such as $f = \text{ERM}(S)$.

So this shows that concentration is not enough.

The recipe is: Given one realization of data set S, find the worst predictor. The metric is based on S, while the predictor might or might not depend on S, just need it to be the worst.